ANALYSIS OF CORRELATION IN A DETERMINATELY CLOSED SYMMETRIC MULTIVARIATE SYSTEM

by

WILLIAM KYRAN WINTERS

B. S., Utah State University, 1960

A MASTER'S THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1963

Approved by:

Major Professor

TABLE OF CONTENTS

INTRODUCTION

In the past, the interpretation of multivariate systems has been dealt with in several ways from a working methodological point of view. First is the inspection of arrays of simple correlation coefficients either at random or in some systematic fashion. Then, from the magnitude of these simple correlation coefficients and knowledge[1] concerning the variables involved, the experimenter or statistician will try to extract what information is needed, perhaps by use of multiple regression or just from the correlations themselves.

Secondly, the interpretation of the multivariate system may be such that prediction, classification, or the regression of one variable on another is the goal of the interpretation. In this case, the methods involved fulfill their aim quite successfully.

Since in many analyses of multivariate systems the interpretation of the system is the information sought because it is the answer to some of the questions concerning the system, the author has considered the following question: Why is multiple regression analysis and inspection of arrays of simple correlation coefficients not adequate for the experimenter in applied areas who has much knowledge in his own area concerning a particular system?

---

[1]This knowledge, which may be of the form of experience in working with various variables or of some other form, can be used to advantage and will be one basic assumption underlying the methods which follow.

One possible answer might be that he has sufficient knowledge of a system in his own field to make any information which the multiple regression or simple correlations may provide just verification of what he already suspects. This, in some cases, may be his goal; but in other instances, may be wasted time and effort for he may also be interested in obtaining some additional information from the data which may be disguised by multivariate interaction. Pursuing this line of thought, it might seem reasonable to ask one more question. Can any additional information be gained from simple correlation coefficients in a multivariate system if knowledge and an assumption toward the system is taken and felt to be true by the experimenter? A partial answer to this question might be given if one were to think in terms of systems of correlation coefficients associated with a multivariate system. Wright (1954) has considered such an approach.

Carrying the idea of a system of correlation coefficients farther, it may be informative to visualize such a system from two angles. First, one may consider the usual descriptive statistical side (Wright 1954). On the other hand, there is the interpretive approach in which one considers causal relations. This method of approach leads to an interpretable solution to the problem.

First, to consider what has already been done along the line of interpretation of a multivariate system and what fruitful ideas can be obtained from the literature, an investigation and short survey of path analysis and the method of path coefficients is warranted. The original work in this area was published by Sewall Wright in 1918 and 1921. Subsequently, it has

been elaborated and re-explained by the same techniques in other works which appeared in 1934, 1954, 1960a, and 1960b. Also, to a very limited extent, others, namely, Tukey (1954); Li (1956); and Turner (1959, 1961), have published what seem to be variations and reinterpretations of the same techniques. Several applications of the techniques have been published by Lu (1959, 1962).

It seems that no one has really presented the interpretation of a system of correlation coefficients from the viewpoint exclusively of an analysis of correlation with a practical solution provided by a computer program. Therefore, it is the intent of the author in this paper to develop, in matrix form, the techniques for such an analysis of correlation in which the underlying multivariate system is symmetric and closed, and to present a practical means for handling high-order multivariate systems.

METHOD

Consider a symmetric system of correlation coefficients denoted by

$$\left[(r_{ij})_s\right] \tag{1}$$

where s is the order of the system and r_{ij} are simple correlation coefficients with $|r_{ij}| \leq 1$ if $i \neq j$ and $r_{ij} = 1$ if $i = j$. Further assume that this system is associated with a determinately closed symmetric multivariate system in which the underlying model has a defined point of view. By a determinately closed system is meant a system of correlation coefficients in which no further variation can be explained by the addition of more factors to the associated multivariate system.

It has been shown by Wright (1960a) that the system in its matrix repre-

sentation is

$$AP = R, \tag{2}$$

where

$$A = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{21} & 1 & r_{23} & \cdots & r_{2n} \\ r_{31} & r_{32} & 1 & \cdots & r_{3n} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ r_{n1} & r_{n2} & r_{n3} & \cdots & 1 \end{bmatrix}, \quad P = \begin{bmatrix} p_{1m} \\ p_{2m} \\ p_{3m} \\ \cdot \\ \cdot \\ \cdot \\ p_{nm} \end{bmatrix}, \quad R = \begin{bmatrix} r_{1m} \\ r_{2m} \\ r_{3m} \\ \cdot \\ \cdot \\ \cdot \\ r_{nm} \end{bmatrix}$$

and m is the effect factor of the system. Note also from (1) that $s = n + 1$. In considering an analysis of correlation of this system, it will be necessary to work with each component of R.

At this point, two terms must be defined, namely, direct effects and compound effects. By a direct effect shall be meant the direct influence of the ith factor to the mth factor of the r_{im} correlation in the system, denoted by p_{im}. A compound effect is the indirect influence of the other factor(s) exclusive of the ith factor of the r_{im} correlation in the system. It should be noted here that the column vector P is, in terms of Wright's (1934, 1960a) work, known as a vector of path coefficients of the determinately closed system. Further, Wright (1934) has shown that the p_{im} measure the directional contribution of the ith factor to the mth factor, since the p_{im} have both magnitude and direction and are independent of physical units, because they are standardized partial regression coefficients. Thus, by pooling the other additive factors, which contribute to r_{im}, exclusive of the p_{im}, a measure of the indirect contribution to

r_{im} is obtained. By definition, the compound effect is an indirect contribution.

Since a symmetric system of correlation coefficients is to be considered and an investigation of the correlation coefficients between the causal factors and the effect factor is desired, it will be advantageous to consider a diagram for each system, which will represent the determinately closed system. Thus, for lack of better means, the convention of Wright (1934, 1960a) is followed in the use of unidirectional and bidirectional arrows supplemented by appropriate correlation and path coefficients as representative of (1). Following this procedure, consider figure 1.

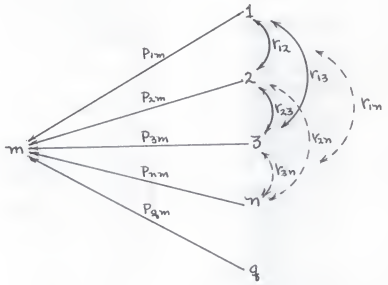


Figure 1. A determinately closed symmetric multivariate system. Some of the features of the preceding diagram are: m represents the effect factor and 1, 2, 3, ..., n represent the causal factors. The q is the residual factor and is assumed to be uncorrelated with the variables 1, ..., n in the closed system. Also, note that the p_{im} (i = 1, ..., n) are represented by directional arrows expressing the directional influence of each p_{im}. The r_{ij} are represented by bidirectional arrows indicating their associative relationship.

For purposes of convenience, the square symmetric matrix A is partitioned into n row vectors in the following manner:

$$A = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{21} & 1 & r_{23} & \cdots & r_{2n} \\ r_{31} & r_{32} & 1 & \cdots & r_{3n} \\ & \vdots & & & \\ r_{n1} & r_{n2} & r_{n3} & \cdots & 1 \end{bmatrix} = \begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ \vdots \\ A_n \end{bmatrix}.$$

Now it can be seen that each $A_i P = R_i$ (i = 1, ..., n), where the R_i are defined by

$$R = \begin{bmatrix} r_{1m} \\ r_{2m} \\ r_{3m} \\ \vdots \\ r_{nm} \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ \vdots \\ R_n \end{bmatrix}$$

is an equation which can be partitioned into a direct and compound effect. Thus

R_i = (Directional effect of the ith factor) + (Compound effect of the other factors).

For example,

$$R_1 = (p_{1m}) + (r_{12}p_{2m} + r_{13}p_{3m} + \cdots + r_{1n}p_{nm})$$
$$= (p_{1m}) + (\sum_{j=2}^{n} r_{1j}p_{jm})$$
$$R_2 = (p_{2m}) + (r_{12}p_{1m} + \sum_{j=3}^{n} r_{2j}p_{jm}).$$

In general, let us represent R_i by

$$R_i = (p_{im}) + (\sum_{j=1}^{n} r_{ij}p_{jm})_{j\neq i}. \qquad (3)$$

Now, for purposes of clarity, consider the following analysis of correlation table:

Table 1 - Analysis of Correlation

Effect	Contribution
Direct	p_{im}
Compound	$\sum_{j=1}^{n} r_{ij}p_{jm} (j\neq i)$
Total correlation	R_i

It is easily seen that table 1 displays, in clear terms, that part of the total correlation which is due to the ith factor directly and also that part which is indirectly responsible for the rest of the total correlation. In a system of order s, there will be s correlation coefficients which may be analyzed. These are represented by the vector R. Further information can be obtained if one wishes to partition the compound effect into individual indirect effects of the s - 1 other factors of the system exclusive of the ith factor. To see this, consider table 2.

Table 2 - Analysis of Correlation

Effect	Contribution
Direct	p_{1m}
Compound	$\sum_{j=1}^{n} r_{1j}p_{jm} \quad (j \neq 1)$
Due to 1	$r_{11}p_{1m}$
Due to 2	$r_{12}p_{2m}$
⋮	⋮
Due to n	$r_{1n}p_{nm}$
Total correlation	R_1

One can easily see the advantage of table 2 since each factor, exclusive of the ith, appears with a magnitude of the contributing or detracting effect, whichever it may be, it has on the total correlation.

A determinately closed system has been under consideration, but there has been no expression of the determination of the closed system by a quantitative measure. It is apparent that the measure should be of such nature that it will be in the range from zero to one. Thus, by multiplying both sides of equation (2) on the left by P', one has

$$P'(AP) = P'R, \tag{4}$$

and by the law of associativity of matrix multiplication

$$(P'A)P = P'R. \tag{5}$$

Letting

$$P'A = B \text{ and } P'R = S$$

the equation becomes

$$BP = S. \tag{6}$$

It may not be obvious, but upon investigation, it can be seen that S is of dimension 1 x 1 since $B_{1 \times n} P_{n \times 1} = S_{1 \times 1}$; thus, it is a scalar and is identical to the coefficient of determination, or the squared multiple correlation coefficient as it is known in multiple regression theory. Now, for the system in (2), there is a measure, in the appropriate range, which is representative of the amount of variation in m which is accounted for by the factors 1, ..., n in the closed system. In addition, it is known that

$$S + (1 - S) = 1 \tag{7}$$

or

$$S + Q = 1 \tag{8}$$

where $Q = 1 - S$. Therefore, if the unexplained variation in the system is of interest, it is seen that Q is such a measure. Thus, it will be convenient to define Q as the coefficient of indetermination of the system. It is interesting to observe that if $Q = 0$, then $S = 1$, and the determinately closed system becomes a completely closed system. If S is something less than one, then the degree to which the system is closed is S x 100%. It is noteworthy also to observe that the coefficient of determination S is the sum of n terms namely

$$\begin{aligned} S &= [r_{1m}p_{1m} + r_{2m}p_{2m} + \ldots + r_{nm}p_{nm}] \\ &= [r_{1m}p_{1m}] + [r_{2m}p_{2m}] + \ldots + [r_{nm}p_{nm}] \end{aligned} \tag{9}$$

where each of the n terms on the right of (9) is representative of the contribution that the ith variable ($i = 1, \ldots, n$) makes to the total determination of the whole system. The significance of this additive breakdown lies in the fact that each variable may be ranked according to its importance in the system.

TABLE 3

		Factor 1	2	3	4
	1	1	r_{12}	r_{13}	r_{14}
	2	r_{21}	1	r_{23}	r_{24}
	3	r_{31}	r_{32}	1	r_{34}
	4	r_{41}	r_{42}	r_{43}	1
Total correlation	5	r_{15}	r_{25}	r_{35}	r_{45}
Direct Effect		p_{15}	p_{25}	p_{35}	p_{45}
Pooled indirect Effect		$\sum_{j=2}^{4} r_{1j}p_{j5}$	$\sum_{j=1}^{4} r_{2j}p_{j5}$ $(j\neq 2)$	$\sum_{j=1}^{4} r_{3j}p_{j5}$ $(j\neq 3)$	$\sum_{j=1}^{3} r_{4j}p_{j5}$
Indirect Effect due to factor	1		$r_{12}p_{25}$	$r_{13}p_{35}$	$r_{14}p_{45}$
	2	$r_{21}p_{15}$		$r_{23}p_{35}$	$r_{24}p_{45}$
	3	$r_{31}p_{15}$	$r_{32}p_{25}$		$r_{34}p_{45}$
	4	$r_{41}p_{15}$	$r_{42}p_{25}$	$r_{43}p_{35}$	
Determination of factor		$r_{15}p_{15}$	$r_{25}p_{25}$	$r_{35}p_{35}$	$r_{45}p_{45}$
Determination of system		$\sum_{j=1}^{5} r_{j5}p_{j5}$			

To summarize the analysis of correlation, consider table 3 for the special case of a system of correlation coefficients of order 5. One can observe that there is one column of the table for each variable and for a system of order less than 8 this is a concise method for summarizing the results. For higher order systems, the computer program considered later, will have an output format which will also be in a somewhat similar form for easy inspection.

NUMERICAL EXAMPLE

A publication which appeared in 1959 by Dewey and Lu is an excellent example of the application of the previous method. This publication dealt with crested wheatgrass seed production which is currently of much interest to the plant breeder. Data were collected on open-pollination progenies of 79 standard crested wheatgrass plants at the Evans farm, which is part of the Utah State University Agricultural Experiment Station, located near Logan, Utah. The authors of the above paper were interested in an investigation of the associations between the characteristics on which measurements were taken for the purpose of determining a selection program to use in order to increase seed production. These characteristics were seed size, spikelets per spike, fertility, plant size, and seed yield.

Table 4 displays the results in which the following points might be brought out as an example of the usefulness of an analysis of correlation in a determinately closed symmetric multivariate system. Upon inspection of the correlations between seed yield and the other four factors which are assumed to determine it, it may be observed that from the correlation coefficients alone, one would suspect that spikelets per spike would be the

TABLE 4

		Factor			
		Seed Size 1	Spikelets per spike 2	Fertility 3	Plant size 4
Seed size	1	1	.274	-.706	.525
Spikelets per spike	2	.274	1	-.300	.474
Fertility	3	-.706	-.300	1	-.665
Plant size	4	.525	.474	-.665	1
Seed yield	5	.005	.477	.256	.440
Direct Effect		.228	.307	1.120	.914
Pooled indirect Effect		-.223	.160	-.864	-.474
Indirect Effect due to factor	1		.062	-.161	.120
	2	.087		-.095	.150
	3	-.790	-.336		-.744
	4	.480	.434	-.608	
Determination of factor		.001	.151	.287	.402

Determination of system .841

most important factor in determining seed yield. This would probably result by observing that .477 > .440 > .256 > .005 which are the correlation coefficients between seed yield and spikelets per plant, plant size, fertility, and seed size respectively. However, by analyzing the vector of correlations (represented by the fifth row in table 4) between seed yield and its determining factors for the direct and indirect effects of each correlation, one's opinion most assuredly would be changed as to the importance of spikelets per spike in determining seed yield when it is seen that the indirect effect of the other factors is the effect which is contributing to the magnitude of the correlation between seed yield and spikelets per spike. Therefore, by considering the direct effects of each factor, it is most apparent that fertility and plant size are the factors to be noted in determining what procedure to use to increase seed yield.

By the use of an analysis of correlation, as appeared in the aforementioned publication, it was possible to suggest those factors in the determinately closed system, namely, fertility and plant size which the plant breeder should use in his selection program.

The usefulness of the method is summarized by Dewey and Lu (1959) in the following manner:

> As more variables are considered in the correlation table, these indirect associations become more complex, less obvious, and somewhat perplexing. At this point, the path-coefficient analysis provides an effective means of untangling direct and indirect causes of association and permits a critical examination of the specific forces acting to produce a given correlation and measures the relative importance of each causal factor.

COMPUTER PROGRAM

Function

The objective of an analysis of correlation program, from a mathematical point of view, is to find the solution to higher order correlation systems of the kind

$$AP = R$$

where A is nonsingular and positive definite by finding the solution vector

$$P = A^{-1}R$$

which involves obtaining the inverse to the correlation matrix A. Then, by suitably partitioning R and A, one can find the direct and indirect coefficients that make up each correlation in the vector R. Thus, the equation becomes

$$A_1P = R_1$$

for all i = 1, 2, 3, ..., n. Then, upon obtaining the solution vector, P' is also determined so that it is now possible to calculate the coefficient of determination

$$S = RP.$$

Mathematical Discussion

Let

$$A = (a_{ij}) = \begin{bmatrix} a_{11} & a_{12} \cdots a_{1n} \\ a_{21} & a_{22} \cdots a_{2n} \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ a_{n1} & a_{n2} \cdots a_{nn} \end{bmatrix}, \text{ where } \begin{cases} |a_{ij}| < 1 \text{ if } i \neq j \\ a_{ij} = 1 \text{ if } i = j. \end{cases}$$

and by augmenting A on the right by the identity matrix I_n,

$$AI_n = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & 1 & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & a_{2n} & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & & \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} & 0 & 0 & \cdots & 1 \end{bmatrix} .$$

By performing elementary row operations of type II, being

$$a'_{ij} = a_{ij}/a_{ii}$$

and of type III being

$$a'_{kj} = a_{kj} - a_{ii}a_{ij}$$

for i, j = 1, ..., n, AI_n can be reduced to the form

$$AI_n \longrightarrow \begin{bmatrix} 1 & 0 & \cdots & 0 & b_{11} & b_{12} & \cdots & b_{1n} \\ 0 & 1 & \cdots & 0 & b_{21} & b_{22} & \cdots & b_{2n} \\ \cdot & \cdot & & \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & 1 & b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix} = I_n B$$

where

$$B = A^{-1} .$$

Now by taking the product

$$BR = P$$

which is the same as $\sum_{i=1}^{n} b_{ji} r_{im}$ for all j = 1, ..., n, the solution to the vector P can be obtained. By considering each ith row separately, one can now analyze each element of the vector R. Thus

$$A_1 P = R_1 .$$

Then by taking the product

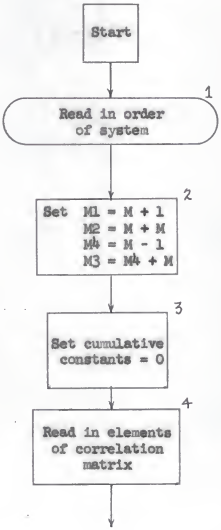
$$P'R = S$$

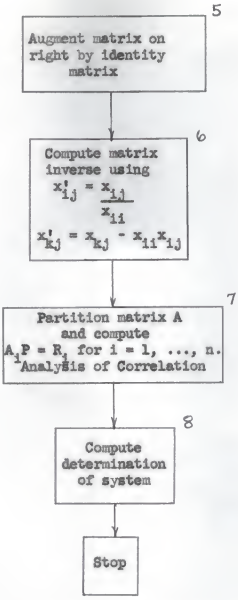
or

$$\sum_{i=1}^{n} p_{im} r_{im} = S$$

the coefficient of determination S is obtained.

Flow Chart





Description of Flow Chart

Box 1: The order M of the system is read into memory.

Box 2: The limits for the summations to be made later are set.

Box 3: A pair of dummy constants are used for the initial starting point of any summation. These are set equal to zero.

Box 4: The elements of the correlation matrix are read and put into memory.

Box 5: An identity matrix of order N is augmented to the right of the correlation matrix.

Box 6: Two elementary matrix operations, namely, type II and type III operations, are used to reduce the correlation matrix A to the identity matrix while at the same time the same operations that were performed on A are performed on the augmented identity matrix to finally reduce it to a matrix A^{-1} which is the inverse of A. Then, the vector P is found by taking the product $A^{-1}R$.

Box 7: Each correlation R_i of the vector R is now found for i = 1, ..., n by taking the product $\sum_{j=1}^{n} r_{ij}p_{jm}$ and for each i and j is printed out.

Box 8: The determination of the system is then computed by taking $\sum_{i=1}^{n} r_{im}p_{im} = S$.

C FORTRAN PROGRAM

```
    1 FORMAT(I4)
    2 FORMAT(F10.6)
   71 FORMAT(24H ANALYSIS OF CORRELATION)
   72 FORMAT(28HCOEFFICIENT OF DETERMINATION)
   80 FORMAT(33H DIRECT CONTRIBUTION COEFFICIENTS)
   81 FORMAT(//)
   83 FORMAT(/)
  100 FORMAT(F18.6,I4,I4)
  310 FORMAT(F10.5,I4,I4)
      DIMENSION S(20,20),B(20),ANAL(20),BAL(20),R(20,20),T(20)
C INITIALIZE AND READ IN DATA
```

```
   51 READ1,M
      M1=M+1
      M2=M+M
      M4=M-1
      M3=M4+M
      DET=0.0
      ANAL(0)=0.0
      DO5I=1,M
      DO5J=1,M
      READ2,S(I,J)
      R(I,J)=S(I,J)
    5 CONTINUE
C AUGMENT MATRIX ON RIGHT BY IDENTITY MATRIX
      DO8I=1,M
      II=I+M
      DO7J=M1,M2
    7 S(I,J)=0.0
    8 S(I,II)=1.0
C COMPUTE DIRECT CONTRIBUTION COEFFICIENTS AND INVERSE,TYPEOUT COEFFS
      PRINT81
      PRINT80
      PRINT81
      DO22I=1,M
   10 FAC=S(I,I)
      IF(I-M)17,11,17
   11 DO13J=1,M4
   12 II=J+M
      B(J)=S(J,M)
   13 PRINT310,B(J),J
      PRINT83
      DO16J=M1,M3
   14 II=J-M
      DO16K=II,M4
   15 TEMP3=S(K,J)
      IF(SENSE SWITCH 3)61,16
   61 PRINT100,TEMP3,K,II
   16 CONTINUE
      PRINT83
   17 DO18J=1,M2
   18 S(I,J)=S(I,J)/FAC
      DO22K=1,M
      IF(K-I)20,22,20
   20 FAC=S(K,I)
      DO21J=1,M2
   21 S(K,J)=S(K,J)-FAC*S(I,J)
   22 CONTINUE
```

```
C PUNCH AND PRINT OUT INVERSES
      DO201K=1,M
      DO200J=M1,M2
      IF(SENSE SWITCH 3)200,201
  200 TYPE100,S(K,J),K,J
  201 CONTINUE
C PRINT OUT ANALYSIS OF CORRELATION
      PRINT83
      PRINT71
      DO301I=1,M4
      BAL(I)=0.0
      DO300J=1,M4
      ANAL(J)=R(I,J)*B(J)
      BAL(I)=BAL(I)+ANAL(J)
  300 PRINT310,ANAL(J),I,J
      PRINT310,BAL(I)
      PRINT83
  301 CONTINUE
      PRINT72
      PRINT83
      DO400J=1,M4
  400 DET=DET+R(M,J)*B(J)
      PRINT2,DET
      PAUSE
      GO TO 51
      END
```

```
DIRECT CONTRIBUTION COEFFICIENTS


   .22973   1
   .31676   2
  1.12189   3
   .91530   4
```

```
ANALYSIS OF CORRELATION
   .22973   1   1
   .08679   1   2
  -.79206   1   3
   .48053   1   4
   .00500
```

```
 .06294   2   1
 .31676   2   2
-.33656   2   3
 .43385   2   4
 .47699

-.16218   3   1
-.09503   3   2
1.12189   3   3
-.60867   3   4
 .25600

 .12060   4   1
 .15014   4   2
-.74606   4   3
 .91530   4   4
 .44000
```

COEFFICIENT OF DETERMINATION

.842187

Discussion of the Program

The fortran program presented is capable of handling a system of correlation coefficients up to order 20; thus, the limits on M are $2 \leq M \leq 20$. The elements of the correlation matrix A are read in by punched card starting with the element in the first row and first column and followed by the succeeding elements in each row as we move down the row vectors in the matrix. Each correlation is punched on a separate card in a ten-digit field along with the identification of its row and column position in the matrix.

By the use of sense switch control, it is possible to have printed out the inverse of the correlation matrix if sense switch number three is on.

The output medium is done by typewriter and is self-explanatory.

SUMMARY

A method has been developed for a very special multivariate system where one factor can be considered and assumed to be determined by n other factors. The correlations are assumed to be known or available but if not, can be computed from sample data. The method can be very useful for higher order systems where the interaction of many factors becomes practically impossible to separate and interpret.

It is very interesting as well as comforting to know that a solution always exists for the system since the correlation matrix A is nonsingular and positive definite. Thus, when one seeks the practical solution to a system $AP = R$, the solution vector P is obtainable since A^{-1} will always exist for such a system.

Like any method which is developed, it is much easier to use and appreciate this method if there is both a practical and usable means for applying it. The computer program presented furnishes the means to this end, and affords, in addition, an available way for verification of the method. The program presented allows one to analyze systems of correlation coefficients up to order 20. With slight alteration, systems up to order 50 could also be handled.

It is quite apparent that the ultimate usefulness of the method depends upon the ingenuity of the user to apply it. The determination of the effect factor is assumed to be unidirectionally, linearly, and completely determined by the other factors of the system and therefore, careful consideration must be used in applying the method. As Tukey (1954) points out, the original

work of Wright (1918, 1921) was an attempt to go from a purely descriptive method of analyzing data to a functional level treatment of data. However, Wright's method is considered by Tukey and others to be probably at a level somewhere between that which they call a tangential level. The method presented is a special case where a determinately closed system is considered so that the assumption of being completely determined is only met when the determination of the system is S = 100%.

ACKNOWLEDGMENTS

The writer is appreciative of his kind relationship with Dr. Stanley Wearden throughout his graduate study. He also extends his deep feeling of indebtedness to his wife for her understanding and confidence and is most grateful for her effort put forth in typing this paper. Furthermore, the writer would like to express his thanks to Dr. K. H. Lu, Head of the Biostatistics Department at the University of Oregon Dental School for his many ideas and suggestions when the writer was under his guidance in 1960.

REFERENCES

(1) Berkson, J.
"Are there Two Regressions?" Journal of the American Statistical Association, 1960. 45: 164.

(2) Cochran, Wm. G.
Sampling Techniques. First edition. New York: John Wiley, 1953.

(3) Dewey, D. R. and K. H. Lu.
"A Correlation and Path-Coefficient Analysis of Components of Crested Wheatgrass Seed Production." Agronomy Journal, 1959. 51: 515-518.

(4) Fisher, R. A.
Statistical Methods for Research Workers. Eleventh edition. London: Oliver and Boyd, 1950.

(5) _____.
Statistical Methods and Scientific Inference. New York: Hafner Publishing Company, 1956.

(6) Graybill, Franklin A.
An Introduction to Linear Statistical Models. New York: McGraw-Hill Book Company, 1961. Vol. I.

(7) Kempthorne, Oscar.
An Introduction to Genetic Statistics. New York: John Wiley, 1957.

(8) Kendall, M. G. and A. Stuart.
The Advanced Theory of Statistics. New York: Hafner Publishing Company, 1958. Vols. I and II.

(9) Li, C. C.
"The Concept of Path Coefficient and Its Impact on Population Genetics." Biometrics, June, 1956. Vol. 12, No. 2.

(10) _____.
Population Genetics. The University of Chicago Press. Chap. 12.

(11) Lu, K. H. and B. S. Savara.
"A Statistical Analysis of Relative Importance in Factors Affecting Weight Changes in Children from Three to Eight Years Old." Human Biology, February, 1962. Vol. 34, No. 1.

(12) Lush, Jay L.
Animal Breeding Plans. Iowa State College Press. 1962.

(13) Mather, K.
Statistical Methods in Biology. New York: Interscience Publishers, 1946.

(14) Ostle, Bernard.
Statistics in Research. Iowa State College Press. 1952.

(15) Rao, C. R.
Advanced Statistical Methods in Biometric Research. New York: John Wiley, 1952.

(16) Stevens, W. L.
"Asymptotic Regression." Biometrics, 1951. 7: 247-67.

(17) Tukey, J. W.
"Causation, Regression, and Path Analysis." Statistics and Mathematics in Biology. Iowa State College Press. 35-66 p.

(18) Turner, Malcolm E. and Charles D. Stevens.
"The Regression Analysis of Causal Paths." Biometrics, June, 1959. Vol. 15, No. 2.

(19) _____, Robert J. Monroe and Henry L. Lucas.
"Generalized Asymptotic Regression and Non-Linear Path Analysis." Biometrics. March, 1961. Vol. 17, No. 1.

(20) Wright, Sewall.
Systems of Mating and Other Papers. Iowa State College Press.

(21) _____.
On the Nature of Size Factors. Genetics. 1918. 3: 367-374.

(22) _____.
"Correlation and Causation." Journal of Agricultural Research, 1921. 20: 557-585.

(23) _____.
"The Method of Path Coefficients." Annals of Mathematical Statistics, 1934. 5: 161-215.

(24) _____.
"The Interpretation of Multivariate Systems." Statistics and Mathematics in Biology. Iowa State College Press. 11-33 p.

(25) _____. "Path Coefficients and Path Regressions: Alternative or Complementary Concepts?" Biometrics, June, 1960a. Vol. 16, No. 2.

(26) _____. "The Treatment of Reciprocal Interaction, with or without Lag, in Path Analysis." Biometrics, September, 1960b. Vol. 16, No. 3.

APPENDIX

APPENDIX

The underlying model of the multivariate system has the form (Wright 1960a)

$$X_m = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_n X_n + a_r X_r \,. \qquad (1')$$

By taking the deviation from the means for each term and dividing each term by σ_m we have

$$\left(\frac{X_m - \bar{X}_m}{\sigma_m}\right) = \frac{a_1\sigma_1}{\sigma_m}\left(\frac{X_1 - \bar{X}_1}{\sigma_1}\right) + \frac{a_2\sigma_2}{\sigma_m}\left(\frac{X_2 - \bar{X}_2}{\sigma_2}\right) + \ldots$$

$$+ \frac{a_n\sigma_n}{\sigma_m}\left(\frac{X_n - \bar{X}_n}{\sigma_n}\right) + \frac{a_r\sigma_r}{\sigma_m}\left(\frac{X_r - \bar{X}_r}{\sigma_r}\right)$$

$$X_m = p_{1m}X_1 + p_{2m}X_2 + \ldots + p_{nm}X_n + p_{rm}X_r \qquad (2')$$

where

$$X_i = \frac{x_i - \bar{x}_i}{\sigma_i} \qquad (i = 1, 2, \ldots, n, r, m)$$

and

$$p_{im} = \frac{a_i\sigma_i}{\sigma_m} \qquad (i = 1, 2, \ldots, n, r).$$

The set of equations for the solution to (2') are

$$p_{1m} = r_{12}p_{2m} + \ldots + r_{1n}p_{nm} = r_{1m}$$

$$r_{21}p_{1m} + p_{2m} + \ldots + r_{2n}p_{nm} = r_{2m} \qquad (3')$$

$$\ldots$$

$$r_{n1}p_{1m} + r_{n2}p_{2m} + \ldots + p_{nm} = r_{nm} \,.$$

By using matrix notation, we can express (3') by

$$AP = R \tag{4'}$$

where

$$A = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1n} \\ r_{21} & 1 & \cdots & r_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ r_{n1} & r_{n2} & \cdots & 1 \end{bmatrix}, \quad P = \begin{bmatrix} p_{1m} \\ p_{2m} \\ \cdot \\ \cdot \\ \cdot \\ p_{nm} \end{bmatrix}, \quad R = \begin{bmatrix} r_{1m} \\ r_{2m} \\ \cdot \\ \cdot \\ \cdot \\ r_{nm} \end{bmatrix}.$$

Noting that by multiplying the first equation in (3') by p_{1m}, the second by p_{2m}, ..., the nth by p_{nm}, one obtains the set of equations

$$\begin{aligned} p_{1m}r_{1m} &= p_{1m}^2 + r_{12}p_{1m}p_{2m} + \cdots + r_{1n}p_{1m}p_{nm} \\ p_{2m}r_{2m} &= r_{21}p_{2m}p_{1m} + p_{2m}^2 + \cdots + r_{2n}p_{1m}p_{nm} \\ &\cdots \\ p_{nm}r_{nm} &= r_{n1}p_{1m}p_{nm} + r_{n2}p_{2m}p_{nm} + \cdots + p_{nm}^2 . \end{aligned} \tag{5'}$$

Now by adding the n equations in (5'), one has

$$\sum_{i=1}^{n} p_{im}r_{im} = \sum_{i=1}^{n} p_{im}^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{n} p_{im}p_{jm}r_{ij}, \quad j > i \tag{6'}$$

but realizing that p_{qm}^2, thus

$$1 = \sum_{i=1}^{n} p_{im}r_{im} + p_{qm}^2 = \sum_{i=1}^{n} p_{im}^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{n} p_{im}p_{jm}r_{ij} + p_{qm}^2, \quad j > i. \tag{7'}$$

This is the same result which Wright (1960a) obtains. But in matrix notation what amounts to the same thing is matrix multiplication on the left of each side of equation (4') by P'.

Thus

$$P'(AP) = P'R$$

but by the associative law for matrix multiplication

$$(P'A)P = P'R$$

so

$$BP = S \tag{8'}$$

where

$$B = P'A \text{ and } S = P'R.$$

But we should notice that

$$\begin{aligned} P'A &= [p_{1m} + p_{2m}r_{21} + \cdots + p_{nm}r_{n1}, \ \cdots, \ p_{1m}r_{1n} + \cdots + p_{nm}] \\ &= [r_{1m}r_{2m} \cdots r_{nm}] \\ &= B \end{aligned}$$

and

$$P'R = p_{1m}r_{1m} + p_{2m}r_{2m} + \cdots + p_{nm}r_{nm}$$

Therefore, corresponding to equation (7'), we have

$$1 = S + (1 - S) = S + Q. \tag{9'}$$

ANALYSIS OF CORRELATION IN A DETERMINATELY CLOSED SYMMETRIC MULTIVARIATE SYSTEM

by

WILLIAM KYRAN WINTERS

B. S., Utah State University, 1960

AN ABSTRACT OF A MASTER'S THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1963

Basically, systems of correlation coefficients can be handled from two points of view. One may consider the purely statistical description of such systems or one may consider the interpretation of causal relations (Wright 1954). The latter has been considered in this thesis for the special case of a determinately closed symmetric multivariate system.

The method is presented in matrix form for ease of handling and a practical solution for higher order systems is provided by a computer program. The matrix equation

$$AP = R$$

is associated with a multivariate model which for lack of a better means is represented by a diagram similar to Wright's (1934, 1960a).

A determinately closed system is defined as well as necessary terms such as, direct and indirect effects, for the purpose of presenting an analysis of correlation. The analysis is presented in tabled form for clarity and convenience, (tables 1 and 2). The determination of the system is then derived and it is found that by partitioning the sum of the terms for the expression of the coefficient of determination that the relative importance of each factor may be accounted for. The entire method is then summarized by the use of table 3.

A numerical example is presented to support the usefulness of such a method, namely, the publication which appeared in 1959 by Dewey and Lu. From this publication, it is quite apparent that in seeking a selection program to use in order to increase seed yield in the crested wheatgrass,

one cannot depend only upon the inspection of correlation coefficients to give an adequate solution to the problem. However, upon utilization of the method of path coefficients, the authors were able to suggest a selection program to follow. It happens that the path analysis model chosen by Dewey and Lu is completely analogous to the method presented in this thesis.

A practical means for the application of the method for higher order systems is furnished by the computer program. Then, presented in detail, is the function of the program, a mathematical discussion, as well as a flow chart with description and a fortran program. The numerical example is then used as a sample output for the computer program.

Some of the main points, such as the basic assumptions underlying the method and when one will have a solution, are then summarized.